Querying astronomical data services in Natural Language

André Schaaff, Thomas Boch, Sébastien Derriere
Centre de Données astronomiques de Strasbourg

Pierre Barjon (ENSIIE Strasbourg), Aymon Deschaint-Acheul (IUT Nancy)

IVOA Santiago
Semantics Session

CDS
CENTRE DE DONNÉES
ASTRONOMIQUES DE STRASBOURG

IVOA

Asterics
Astronomy ESFRI & Research Infrastructure Cluster

□ Purpose

- The aim is to relate a budding R&D work done in the frame of Natural Language Processing applied to the querying of astronomical data services and to collect comments (bad or enthusiastic), ideas, recommendations, or to initiate collaborations

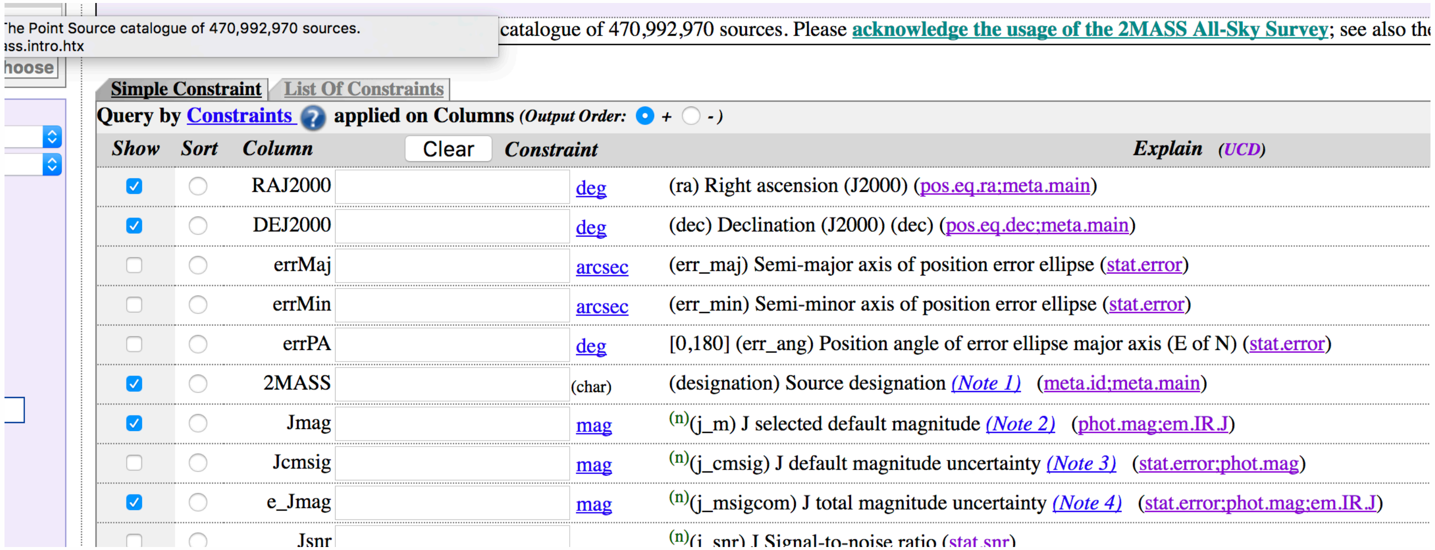
- Are we only geeks ?

- Are we on the way, modestly, to the future (and common) interaction between human and devices ?

- … and is it possible to reach query results satisfying professional astronomers ?

□ Are we only geeks ? (no !)

- Voice usage is becoming natural (Siri, Ok Google, etc.)
- An alternative to the current way based on forms (parameter fields, checkboxes, etc.) through a unique text field or a voice recognition of its content



he Point Source catalogue of 470,992,970 sources.
ss.intro.htx

catalogue of 470,992,970 sources. Please acknowledge the usage of the 2MASS All-Sky Survey; see also th

hoose

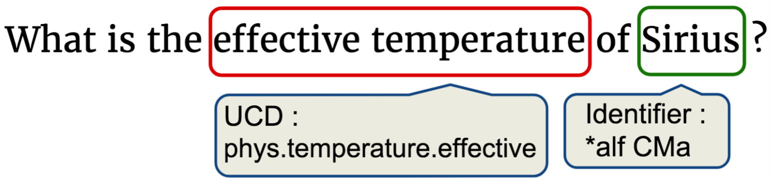
Simple Constraint | List Of Constraints

Query by Constraints ? applied on Columns (Output Order: ● + ○ -)

Show	Sort	Column	Clear	Constraint	Explain (UCD)
☑	○	RAJ2000		deg	(ra) Right ascension (J2000) (pos.eq.ra;meta.main)
☑	○	DEJ2000		deg	(dec) Declination (J2000) (dec) (pos.eq.dec;meta.main)
☐	○	errMaj		arcsec	(err_maj) Semi-major axis of position error ellipse (stat.error)
☐	○	errMin		arcsec	(err_min) Semi-minor axis of position error ellipse (stat.error)
☐	○	errPA		deg	[0,180] (err_ang) Position angle of error ellipse major axis (E of N) (stat.error)
☑	○	2MASS		(char)	(designation) Source designation *(Note 1)* (meta.id;meta.main)
☑	○	Jmag		mag	$^{(n)}$(j_m) J selected default magnitude *(Note 2)* (phot.mag;em.IR.J)
☐	○	Jcmsig		mag	$^{(n)}$(j_cmsig) J default magnitude uncertainty *(Note 3)* (stat.error;phot.mag)
☑	○	e_Jmag		mag	$^{(n)}$(j_msigcom) J total magnitude uncertainty *(Note 4)* (stat.error;phot.mag;em.IR.J)
☐	○	Jsnr			$^{(n)}$(j_snr) J Signal-to-noise ratio (stat.snr)

How ?

- Learn about NLP (basis, tools, examples, …)
- Define the scope of the study
  - Too large -> too much time and resources
  - A first set of queries
- => a pragmatic approach (more R&D than R)
- We start not from scratch
  - Authors in Simbad, VizieR
  - Missions and wavelengths inVizieR
  - We have DJIN to recognize identifiers in a text
  - We have UCDs
  - We have a name resolver
  - We have ADQL / TAP
  - …



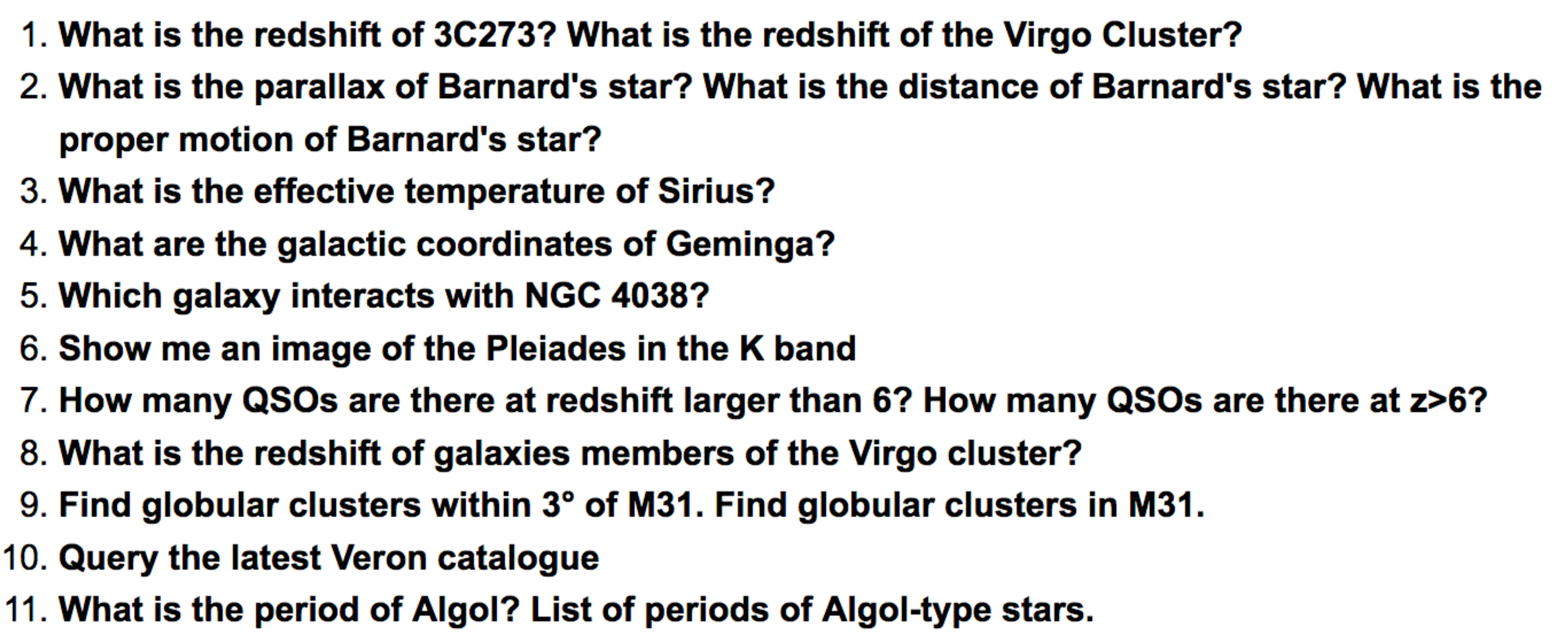
What is the effective temperature of Sirius ?

UCD : phys.temperature.effective

Identifier : *alf CMa

□ Queries in NL

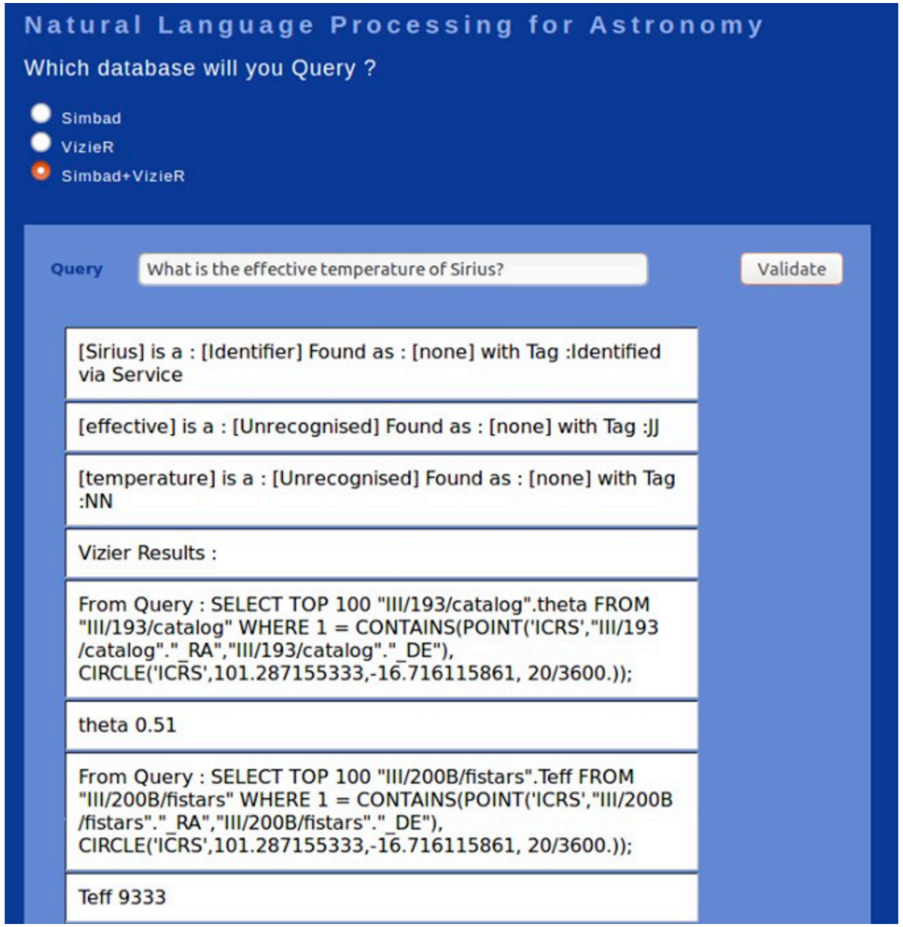
- A wide variety…



1. What is the redshift of 3C273? What is the redshift of the Virgo Cluster?
2. What is the parallax of Barnard's star? What is the distance of Barnard's star? What is the proper motion of Barnard's star?
3. What is the effective temperature of Sirius?
4. What are the galactic coordinates of Geminga?
5. Which galaxy interacts with NGC 4038?
6. Show me an image of the Pleiades in the K band
7. How many QSOs are there at redshift larger than 6? How many QSOs are there at z>6?
8. What is the redshift of galaxies members of the Virgo cluster?
9. Find globular clusters within 3° of M31. Find globular clusters in M31.
10. Query the latest Veron catalogue
11. What is the period of Algol? List of periods of Algol-type stars.

Natural Language Processing to request astronomical services

- A first prototype based on Stanford NLP (POSTagger), DJIN, IVOA UCD and ADQL/TAP, …



Natural Language Processing for Astronomy

Which database will you Query ?

○ Simbad
○ VizieR
● Simbad+VizieR

Query: What is the effective temperature of Sirius?

[Sirius] is a : [Identifier] Found as : [none] with Tag :Identified via Service

[effective] is a : [Unrecognised] Found as : [none] with Tag :JJ

[temperature] is a : [Unrecognised] Found as : [none] with Tag :NN

Vizier Results :

```
From Query : SELECT TOP 100 "III/193/catalog".theta FROM
"III/193/catalog" WHERE 1 = CONTAINS(POINT('ICRS',"III/193
/catalog"."_RA","III/193/catalog"."_DE"),
CIRCLE('ICRS',101.287155333,-16.716115861, 20/3600.));
```

theta 0.51

```
From Query : SELECT TOP 100 "III/200B/fistars".Teff FROM
"III/200B/fistars" WHERE 1 = CONTAINS(POINT('ICRS',"III/200B
/fistars"."_RA","III/200B/fistars"."_DE"),
CIRCLE('ICRS',101.287155333,-16.716115861, 20/3600.));
```

Teff 9333

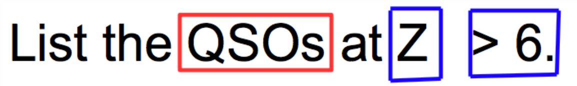
What is the effective temperature of Sirius ?

How many planets orbit Kepler 20 ?

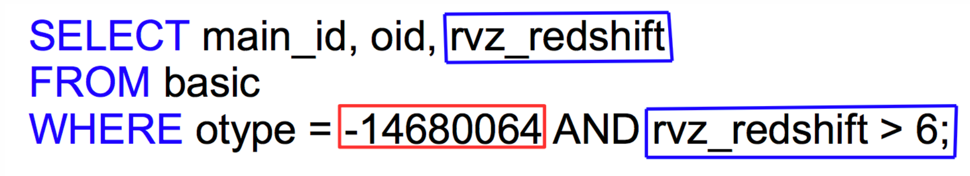
What is the redshift of galaxies members of the Virgo cluster ?

From Natural Language to ADQL

- Examples

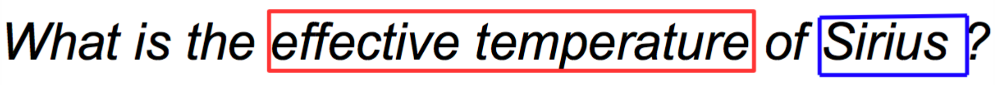


List the QSOs at Z > 6.

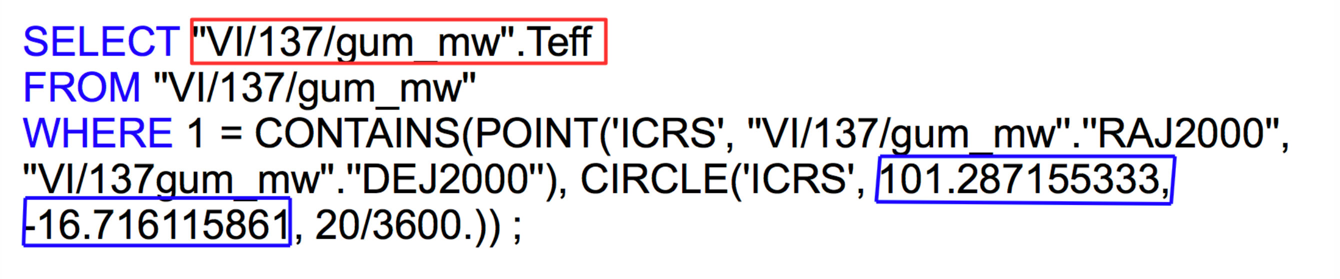


```
SELECT main_id, oid, rvz_redshift
FROM basic
WHERE otype = -14680064 AND rvz_redshift > 6;
```

Simbad, TAP query



What is the effective temperature of Sirius ?



```
SELECT "VI/137/gum_mw".Teff
FROM "VI/137/gum_mw"
WHERE 1 = CONTAINS(POINT('ICRS', "VI/137/gum_mw"."RAJ2000",
"VI/137gum_mw"."DEJ2000"), CIRCLE('ICRS', 101.287155333,
-16.716115861, 20/3600.)) ;
```

VizieR, TAP Query

Comments…

□ Next steps

- Robustness for the first set of queries
- Enlarge this set
- Chatbot approach to fill the gap between “good” queries and unprecise / ambiguous queries

- => realistic and bottom-up approach, a way to create smart portals

□ Please, give us a feedback

- Comments ?
  - bad ?
  - Enthusiastic ?
- Ideas ?
- Recommendations ?
- Collaborations ?

□ Conclusion

- Are we on the way, modestly, to the future (and common) interaction between human and devices ?
- Probably

who knows what one can expect in the future

- And in any case we are not expecting an exhaustive recognition of the queries, it will be improved step by step